Effective Analysis and Predictive Model of Stroke Disease using Classification Methods

A.Sudha
Student, M.Tech (CSE)
VIT University
Vellore, India

P.Gayathri
Assistant Professor
VIT University
Vellore, India

N.Jaisankar
Professor
VIT University
Vellore, India

ABSTRACT
In today's world data mining plays a vital role for prediction of diseases in medical industry. Stroke is a lifethreatning disease that has been ranked third leading cause of death in states and in developing countries. The stroke is a leading cause of serious, long term disability in US. The time taken to recover from stroke disease depends on patients' severity. Number of work has been carried out for predicting various diseases by comparing the performance of predictive data mining. Here the classification algorithms like Decision Tree, Naive Bayes and Neural Network is used for predicting the presence of stroke disease with related number of attributes. In our work, principle component analysis algorithm is used for reducing the dimensions and it determines the attributes involving more towards the prediction of stroke disease and predicts whether the patient is suffering from stroke disease or not.

General Terms
Data mining, Classification algorithm, Stroke disease.

Keywords
Data mining, Classification algorithm, life threatening diseases.

1. INTRODUCTION
The Stroke is a major leading cause of death and has a serious long term disability. If the patient suffers from heart-related disease, coronary sclerosis, ventricular fibrillation arrhythmia, then the condition is easily complicated by Stroke or Cerebrovascular diseases [1]. Hypertension and Hyperlipidemia are the common risk factors of stroke. Stroke is a major life threatening disease which injures the brain, like heart attack which injures the heart. It does not circulate blood and enough oxygen to the brain cells. Everyone overcomes with some stroke risk [18]. The stroke may cause paralysis, sudden pain in chest, speech impairment, loss of memory and reasoning ability, coma, or death. Stroke affects the person on any age. It can be prevented through the therapeutic manipulation and modifiable risk factors are crucial. The recovery may depend on patients' severity. The report mentions the most common medical error occurs due to expiry of medicines, incorrect drugs, incorrect dosages and treatment given to the wrong patient [7, 16]. In this paper predictive classification algorithm like Decision tree, Naïve bayes and neural networks is used for predicting stroke diseases and principle component analysis algorithm for reducing the attributes.

2. METHODOLOGY
The Methodology described in this paper is Prediction of stroke disease which focuses on creating awareness to the people, and studies are based on recent trends and journals from various publications.

2.1 RELATED STUDIES ON PREDICITION OF LIFE THREATNING DISEASES
A predictive model for cerebrovascular disease using data mining [2] utilizes the data mining techniques: and adopts the classification algorithm like decision trees, Bayesian classifier and back propagation neural network. It consists of attributes containing patient's medical history and symptoms. The records with irrelevant data are removed from data warehouse before mining process occurs. Decision tree performance is very high when compared with other two algorithms in predicting cerebrovascular disease. Cerebrovascular disease has been ranked the second or third of top 10 death causes in Taiwan and has caused about 13,000 people death every year since 1986. It is hard to make accurate diagnosis in advance. This model used 493 valid samples with 8 important attributes of patient's data. These attributes includes their physical exam results, blood test results and diagnoses. This study used sensitivity and accuracy indicators for evaluation. Decision tree achieves 95.29% of sensitivity and 98.01% of accuracy. Bayesian classifier achieves 87.10% and 91.30% respectively; BPNN achieves 94.82% and 97.87% respectively. After comparing with the more stable classification efficiency, decision tree was chosen as the best classification algorithm in this study for predicting cerebrovascular diseases.

Artificial Neural Networks (ANN) is used for the prediction of Thrombo-embolic stroke disease [7].This study uses the dataset of patients who have the symptoms of stroke disease. Backward stepwise method is used for input feature selection. Performance of neural network is achieved by removal of insignificant inputs. This research work demonstrates about ANN based prediction of stroke disease by improving the accuracy to 89% with higher consistent rate. The ANN exhibits good performance level for prediction of stroke disease.

The data mining methods like artificial neural network technique is used in effective heart attack prediction system. First the dataset used for prediction of heart diseases was pre-processed and clustered by means of K-means clustering algorithm [6]. Then neural network is trained with the selected significant patterns. Multi-layer Perceptron Neural Network with Back-propagation is used for training. The results indicate that the algorithm used is capable of predicting the heart diseases more efficiently. The prediction of heart diseases significantly uses 15 attributes, with basic data mining technique like ANN, Clustering and Association Rules, soft computing approaches etc. The outcome shows that Decision Tree performance is more and some times

Bayesian classification is having similar accuracy as of decision tree but other predictive methods like K-Nearest Neighbor, Neural Networks, Classification based on clustering will not perform well [5]. By using the Weighted Associative Classifier (WAC), a slight change has been made, instead of considering 5 class labels, only 2 class labels are used. One for "Heart Disease" and another one for "No Heart Disease". The maximum accuracy (81.51%) has been achieved[15]. When genetic algorithm is applied, the accuracy of the Decision Tree and Bayesian Classification is improved by reducing the actual data size. The dataset of 909 patient records with heart diseases has been collected and 13 attributes has been used for consistency [5]. The patient records have been splitted equally as 455 records for training dataset and 454 records for testing dataset. After applying genetic algorithm the attributes has been reduced to 6 and decision tree performs more efficiently with 99.2% accuracy when compared with other algorithms.

3. DATA MINING TECHNOLOGY

Data mining classification technology consists of two models (classification model and evaluation model). The classification model makes use of training data set in order to build classification predictive model. Testing data set is used for testing the classification efficiency. Patient dataset is collected from healthcare institute who have symptoms of stroke disease. Then classification algorithm like decision tree, naive bayes and neural network is used for prediction to find whether patient is suffering from stroke disease with indicating levels shown in (fig 1). Then performance evaluation is carried out based on three algorithms and compared with various models used and accuracy is measured. Then it is compared with existing model and validated, how the proposed model is better than existing models.

Patient Dataset		
Name	**Symptom**	**Years**
Ani	Headache	1
Bill	Loss of weight	2
Mike	Hypertension	< 1
Jim	Diabetes	2
Dave	Hepatitis	2
Elman	Exercise	Regular

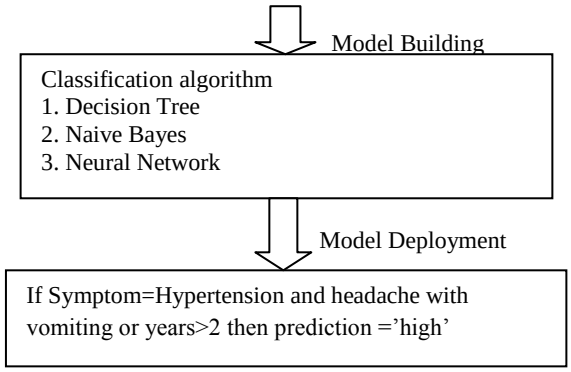

Fig 1: Parameters with risk level

3.1 Decision tree

Decision tree is one of the important method for handling high dimensional data. It looks like a tree structure. It is very simple and easy way for handling dataset. Much work has been carried out to predict the life threatening diseases using decision tree and proved to be more efficient. Fig 2 represents the decision tree model for predicting stroke diseases.

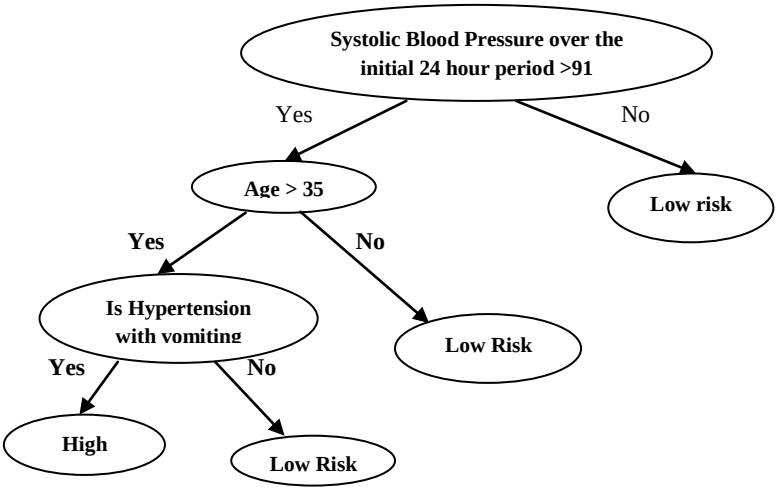

Fig 2: Decision Tree

3.2 Bayesian classifier

A naïve Bayesian classifier depends on Bayes' theorem which works on probabilistic statistical classifier. The major advantage of using this naïve Bayesian classifier includes rapidity of use and very simple for handling the dataset containing many attributes. Fig 3 represents the Naïve bayes model for predicting stroke diseases.

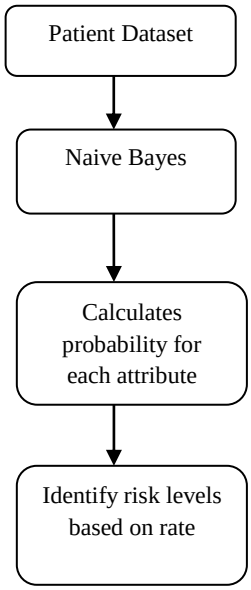

Fig 3: Bayesian classifier

3.3 Neural Network

Neural Network (NN) is a collection of neurons which has connectivity between two or more network layers. It is made up of input layer, zero or more hidden layers and output layer. The input layer has patient dataset as inputs into the network units, connected to a layer of hidden units. The hidden units, is then interconnected to a layer of output units. Fig 4 represents the Neural Network model for predicting stroke diseases.

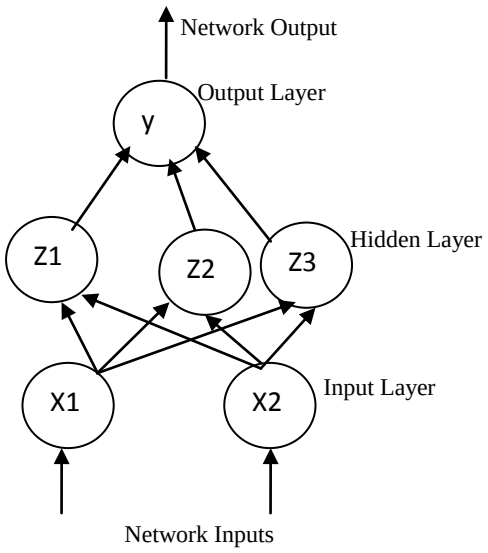

Fig 4: Neural Network

4. PROPOSED MODEL

The proposed model uses predictive classification algorithms like Decision tree, Naïve bayes and neural networks for predicting the presence of stroke disease and Principle Component Analysis algorithm for dimensionality reduction. So the reduced subset of the attributes could be used as inputs. The Fig -5 represents the Prediction of Stroke Disease.

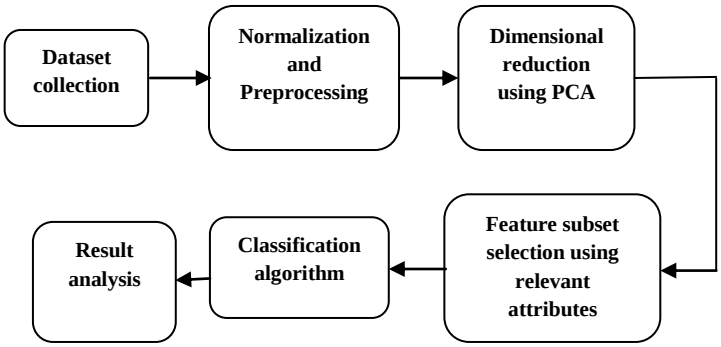

Fig 5: Architecture diagram for prediction of Stroke disease.

4.1 Dataset Collection

In the first step the stroke dataset is collected from the medical institute. The dataset consists of patient information, patient history, Gene diagnosis disease database which contains the symptoms of stroke disease. The data are built to be error free in nature. All the symptoms are analyzed carefully for the prediction of stroke. Analysis of gene expression data leads to stroke identification and classification in earlier stage.

4.2 Pre-processing Dataset

Initially, the stroke disease dataset is pre-processed to make it suitable for mining process. The Pre-processing technique removes the duplicate records, missing data, noisy and inconsistent data.

4.3 Dimensional Reduction using PCA algorithm and Clustering

Once pre-processing is over, the dataset containing more than 1000 records is used for predicting stroke diseases. By using these records it is very difficult and time consuming task to identify the diseases. Hence Principle component analysis is used. It deals with huge amount of dataset and reduces it to a lower dimension. Then clustering is used to group those similar datasets. The attribute values need to be normalized before clustering, in order to avoid the high value attributes that may confuse or underestimate the low value attributes. The PCA algorithm steps are as follows:

1. The mean value for the dataset is calculated and produces the data set whose mean is zero.
2. Then calculate covariance matrix:
$$C^{m\times n} = (c_{i,j}, c_{i,j} = cov(Dim_i, Dim_j))$$
3. Calculate the eigenvectors and eigenvalues of the covariance matrix.
4. From the list of eigenvectors take the eigenvectors selected and form a matrix with them in the columns:
   FeatureVector = (eig_1, eig_2, ..., eig_n)
5. By taking the transpose of the Feature Vector and multiply it on the left of the original data set, transposed:

FinalData= RowFeatureVector x RowDataAdjusted

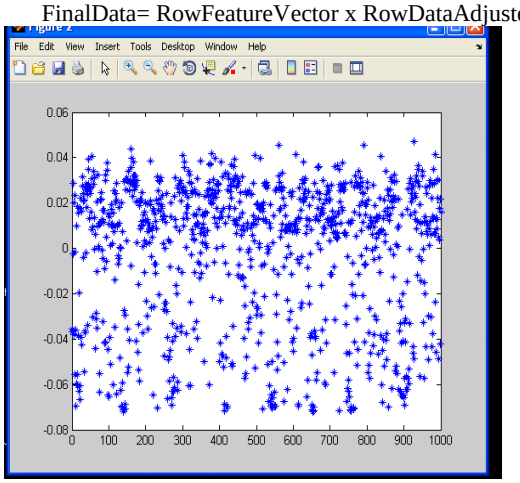

Fig 6: Before Reduction

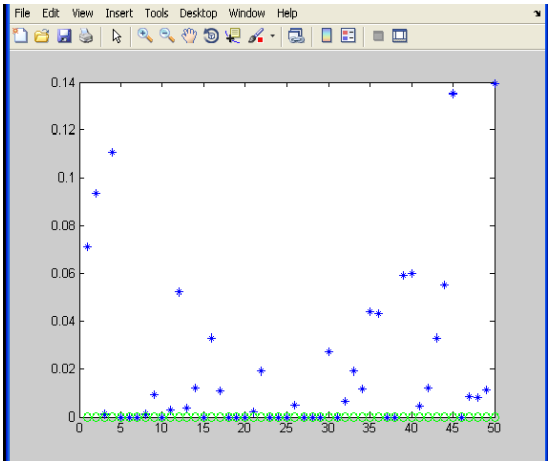

Fig 7: After Reduction

4.4 Feature subset selection using relevant attributes

Feature subset selection is mainly used for feature reduction. It removes the irrelevant data and selects the data which are related to stroke disease. In order to predict the stroke disease it adopts classification algorithms like Decision tree, Bayesian classifier and neural network. These algorithms take training expression data set as input and predict the stroke disease with various levels (normal, medium or high). Thus, Classification algorithms are applied for training and testing data sets and their results are evaluated to determine the most significant gene set. In the first stage the stroke training gene data set is initially separated into several subsets with approximate genes in each subset. It consists of two parts, construction of classification model and evaluation model for finding classification efficiency. The comparison is done with these algorithms and the result indicates by reducing minimum number of attributes and showing accuracy.

Example

```
for i=1:L
   if  Blood Pressure(i)<=119/79
      B(i)=0; %Low
   Elseif BP(i)=130/89
      B(i)=1; %Normal
%     elseif BP(i)>200/160;
      B(i)=2;% High
   end
end
Bin1=dec2bin(B);
```

4.5 Classification algorithms

Decision Tree is constructed from a given set of attributes. It starts from root node and the condition is applied based on values. With the outcome of test, it will lead to another internal node and predict the risk levels. Neural network is connectivity between input units, hidden and output units. It is based on weights assigned to the units in various layers. Naive Bayes works on probability statistical classifier for every possible value in the target range. If the patient detail is entered it checks with the reduced attribute (Ex: glucose level, blood pressure level, family history etc...) and find out the patient have stroke disease or not.

5. RESULT ANALYSIS

The accuracy is measured based on sensitivity and specificity. Based on patient risk levels with diagnosis result = normal and diagnosis result = High, number of attributes used with patient suffering from diseases is measured. The experiment is conducted in mat lab. Fig 8 and 9 shows the result analysis and accuracy for predicting stroke diseases.

Sensitivity = t-pos/pos

t_pos is the number of true positives (healthy sample correctly analyzed) and pos is the number of positive samples.

Sensitivity = t-neg/neg

t_neg is the number of true negatives (sick samples correctly analyzed) and neg is the number of negative samples.

$$\textbf{Accuracy=sensitivity}\left(\frac{\textbf{pos}}{\textbf{Pos+neg}}\right)\textbf{+ sensitivity}\left(\frac{\textbf{neg}}{\textbf{pos+neg}}\right)$$

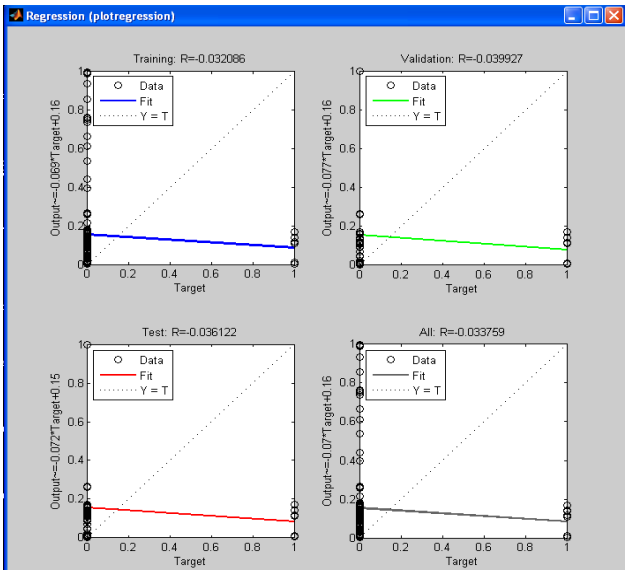

Fig 8: Training and Testing stage

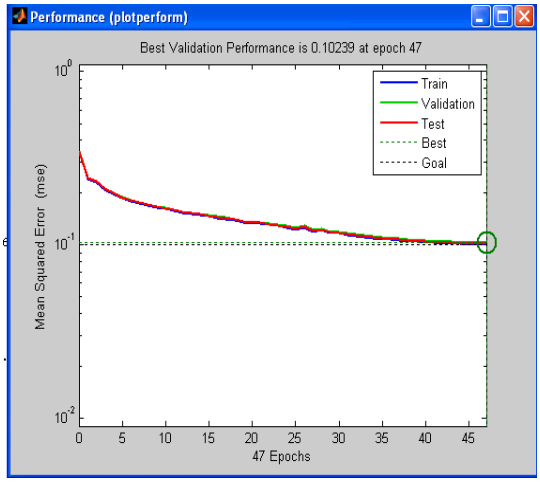

Fig 9: Validation Performance

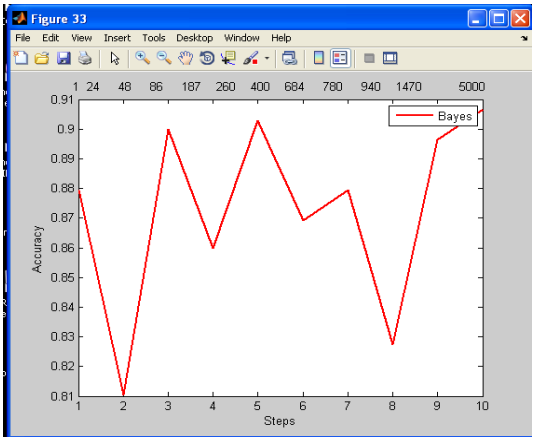

Fig 9.1: Accuracy of Naïve Bayes

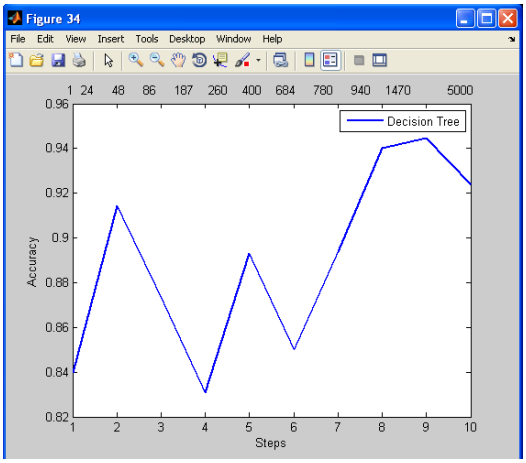

Fig 9.2: Accuracy of Decision tree

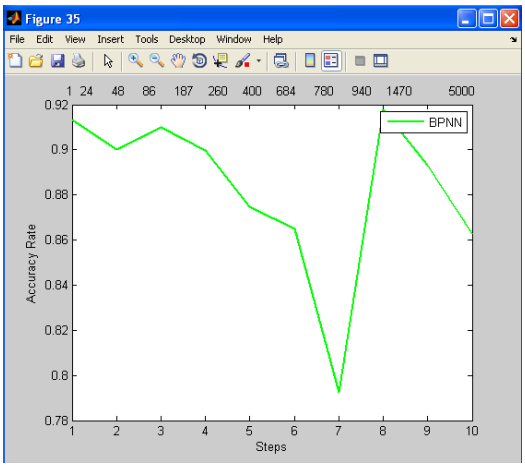

Fig 9.3: Accuracy of Neural network

6. CONCLUSION

Healthcare industry makes use of data mining techniques and generates huge amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices etc. Three classifiers such as decision tree, naive bayes and neural network were used for diagnosis of patients with stroke disease. Observation shows that neural network performance is having more accuracy, when compared with other two classification methods.

7. ACKNOWLEDGMENTS

I thank Dr. N. Jaisankar and Prof. P. Gayathri, VIT University for involving me into this work.

8. REFERENCES

[1] Kohn, L. T., Corrigan, J. M., and Donaldson, M. S., to err is human: building a safer health system. Institute of Medicine (IOM). National Academies Press, Washington, 1999.

[2] Duen-Yian Yeh a, Ching-Hsue Cheng b, Yen-Wen Chen b A predictive model for cerebrovascular disease using data mining‘Science, Vol. 8970-8977, 2011.

[3] Cheng-Ding Chang a, Chien-Chih Wang b, Bernard C. Jiang Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors Vol 38 ,5507–5513, 2011.

[4] Genetics and Genomics of Stroke Novel Approaches Alison E. Baird, MBBS, PHD *Brooklyn, New York* Vol. 56, No. 4, 2010.

[5] M. Anbarasi et. al. Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm International Journal of Engineering Science and Technology Vol. 2(10), 5370-5376 ,2010.

[6] ShantakumarB.Patil,Y.S.Kumaraswamy. ‘Predictive data mining for medical diagnosis of heart disease prediction’jyoti soni, ujma ansari, dipeshsharma IJCSE Vol .17, 2011 .

[7] D.Shanthi,,Dr.G.Sahoo,,Dr.N.Saravanan,2008 ‘Designing an Artificial Neural Network Model for the Prediction of Thrombo-embolic Stroke (IJBB), Volume 3. pp.10-18.

[8] Tamer Uçar a, Adem Karahocaa ‘Predicting existence of Mycobacterium tuberculosis on patients using data mining approaches’ Vol .3, 2011.

[9] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.

[10] Kaur, H., Wasan, S. K.: "Empirical Study on Applications of Data Mining Techniques in Healthcare", Journal of Computer Science 2(2), 194-200, 2006.

[11] G.Subbalakshmi et al. Decision Support in Heart Disease Prediction System using Naive Bayes / Indian Journal of Computer Science and Engineering (IJCSE) Vol. 2 No. 2, 2011.

[12] Shantakumar B.Patil Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network ISSN 1450-216X Vol.31 No.4 pp.642-656, 2009.

[13] Shantakumar B.Patil and Y.S.Kumaraswamy Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450- 216X Vol.31 No.4, pp. 642-656, 2009.

[14] American Heart Association. *Heart Disease and Stroke Statistics — 2004 Update*. Dallas, Tex.: American Heart Association; 2003.

[15] P.K. Anooj Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules Journal of King Saud University – Computer and Information Sciences 24, 27–40, 2011.

[16] A.sudha, P Gayathri and N Jaisankar. Utilization of Data mining Approaches for Prediction of Life Threatening Diseases Survivability. *International Journal of Computer Applications* 41(17):51-55, March 2012.